BEA: Revisiting anchor-based object detection DNN using Budding Ensemble Architecture

Syed Sha Qutub[1,2]
syed.qutub@intel.com
Neslihan Kose[1]
neslihan.kose.cihangir@intel.com
Rafael Rosales[1]
rafael.rosales@intel.com
Michael Paulitsch[1]
michael.paulitsch@intel.com
Korbinian Hagn[1]
Korbinian.Hagn@intel.com
Florian Geissler[1]
florian.geissler@intel.com
Yang Peng[1]
yang.r.peng@intel.com
Gereon Hinz[2]
gereon.hinz@tum.de
Alois Knoll[2]
knoll@in.tum.de

[1] Intel Labs
Munich, Germany

[2] Technical University of Munich
Munich, Germany

Abstract

This paper introduces the Budding Ensemble Architecture (BEA), a novel reduced ensemble architecture for anchor-based object detection models. Object detection models are crucial in vision-based tasks, particularly in autonomous systems. They should provide precise bounding box detections while also calibrating their predicted confidence scores, leading to higher-quality uncertainty estimates. However, current models may make erroneous decisions due to false positives receiving high scores or true positives being discarded due to low scores. BEA aims to address these issues. The proposed loss functions in BEA improve the confidence score calibration and lower the uncertainty error, which results in a better distinction of true and false positives and, eventually, higher accuracy of the object detection models. Both Base-YOLOv3 and SSD models were enhanced using the BEA method and its proposed loss functions. The BEA on Base-YOLOv3 trained on the KITTI dataset results in a 6% and 3.7% increase in mAP and AP50, respectively. Utilizing a well-balanced uncertainty estimation threshold to discard samples in real-time even leads to a 9.6% higher AP50 than its base model. This is attributed to a 40% increase in the area under the AP50-based retention curve used to measure the quality of calibration of confidence scores. Furthermore, BEA-YOLOV3 trained on KITTI provides superior out-of-distribution detection on Citypersons, BDD100K, and COCO datasets compared to the ensembles and vanilla models of YOLOv3 and Gaussian-YOLOv3.

1 Introduction

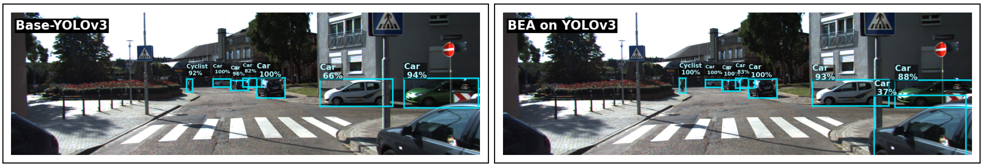


Figure 1: The figure on the left demonstrates uncalibrated confidence scores of Base-YOLOv3, where the white car has a 66% confidence score despite negligible occlusion, and the green car on the right edge has a 94% confidence score despite occlusion. In contrast, the right figure shows that BEA-YOLOv3 detects a car on the bottom right edge, which Base-YOLOv3 misses, and has superior confidence score calibration, indicating the effectiveness of our approach over baseline (Base-YOLOv3).

Object detection deep neural networks (DNN) are a critical technology in numerous fields, such as medical imaging, robotics, and autonomous vehicles. The current state-of-the-art in object detection models can be widely classified into single-stage (YOLOv3 [37], Retina-Net [27], SSD [29]) and two-stage detectors (Faster r-cnn [38], Fast-r-cnn [15], Mask-RCNN [17]). Two-stage detectors are the current state-of-the-art for object detection, as they propose regions of interest before predicting object class and bounding boxes. Single-stage detectors, which directly predict object class and bounding boxes from anchor or prior boxes, are faster but less accurate. Anchor-based single-stage detectors are suitable for real-time applications due to their lower computational overhead. To improve the accuracy of object detection models, researchers have proposed ensemble modelling [32] and post-hoc calibration [16] techniques. There are different ensemble-based approaches, and they generally perform better than a single standalone model, mainly for the following reasons. It captures uncertainty when models are trained with subsets of extensive data. They are more robust to over-fitting as the predictions are averaged, indirectly improving the generalization towards out-of-distribution data. Similarly, the post-hoc calibration techniques are the methods to re-calibrate the predicted probabilities of a model post-training. They are better than no calibration of a model as the confidence scores cannot be directly trusted, but it deals with its inherent issues and is limited by the model's training method. It is also highly biased towards the dataset it is trained with. In [33], the authors claim that post-hoc methods fail to provide well-calibrated uncertainty estimates under distributional shifts in the real world.

This work aims to enhance the quality of the prediction confidence score through calibration (as shown in Figure 1). Our approach builds upon the work of Kornblith *et al.* [18], who demonstrated that wider networks with similar architecture tend to learn more similar feature representations. By measuring multiple similarity scores, they found that the similarity of early layers in these networks plateaus at fewer channels than later layers and that early layers learn comparable representations across diverse datasets. By leveraging these insights, we propose a more efficient and effective reduced ensemble architecture called the **Budding Ensemble Architecture (BEA)**, which includes a common backbone and only two replicated detectors. The name BEA reflects that its two detectors branch out from the backbone. We extensively evaluate our proposed method and demonstrate its superior performance to the current state-of-the-art through a series of experiments on the KITTI dataset [14]. The BEA is equipped with a novel function: **Tandem** loss to enhance calibration and capture out-of-distribution uncertainty in the predictions. The proposed approach results in well-calibrated predictions leading to better uncertainty estimates and aiding in detecting out-of-distribution (OOD) images. Furthermore, we show the quality of the predictions using average precision metric and the quality of our uncertainty measure using uncertainty error (UE) [32] and Retention curves [31].

In summary, this work makes the following contributions:

1. This study presents the Budding Ensemble Architecture (BEA), a novel architecture that outperforms state-of-the-art models in terms of accuracy. The experimental results demonstrate that the BEA enhances confidence score calibration more than the base and ensemble models.

2. We introduce AP50-based retention curves to measure the calibration quality for object detection models.

3. The BEA demonstrates promising results in capturing more accurate detection of out-of-distribution images compared to their respective vanilla and ensemble models.

2 Related Work

Object detection is a complex, high-dimensional task that involves detecting objects within an image and accurately determining their location and size using confidence scores and regression techniques. This makes uncertainty quantification challenging, as many potential error sources must be considered. There are two main approaches to uncertainty estimation in neural network predictions: Bayesian [1, 10, 11, 42, 46] and non-Bayesian [23, 28, 41]. Bayesian approaches use probabilistic models to address uncertainty but are computationally demanding, while non-Bayesian methods use heuristics or techniques to estimate uncertainty, resulting in lower accuracy. Proposed methods for well-calibrated uncertainty estimation primarily focus on classification tasks or are based on post-hoc calibration. In [20], the authors target classification tasks and introduce differentiable accuracy versus uncertainty calibration (AvUC) loss to provide well-calibrated uncertainties and improved accuracy. In [19], the authors propose a loss-calibrated inference method for time-series-based prediction and regression tasks following the insights from [20]. Post-hoc methods [21, 22, 30] also mainly target classification or small-scale regression problems. These methods are based on recalibrating a pre-trained model using a sample of independent and identically distributed (i.i.d.) data. In [21], the authors propose a post-hoc method to calibrate the output of 1D regression algorithms based on Platt scaling [34]. The authors in [40] explored the calibration properties of object detection networks and examined the influence of position and shape in object detection on the calibration properties. Gaussian-YOLOv3 [6] introduces Gaussian modelling of bounding box predictions, where variances are utilized as prediction uncertainty. However, evaluating its performance in terms of confidence score calibration is inadequate. To the best of our knowledge, there are limited numbers of existing calibration solutions for object detection as they bring additional challenges when designing these methods. Ensemble deep neural networks (DNNs) improve DNN model performance and

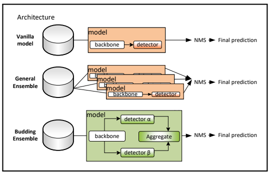


Figure 2: Existing architectures vs Budding Ensemble architecture

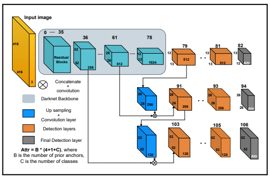

Input image
Residual Blocks
Concatenate + convolution
Darknet Backbone
Up sampling + Convolution layer
Detection layers
Final Detection layer
Attr = B * (4+1+C), where
B is the number of prior anchors,
C is the number of classes

Figure 3: Base model: YOLOv3 architecture

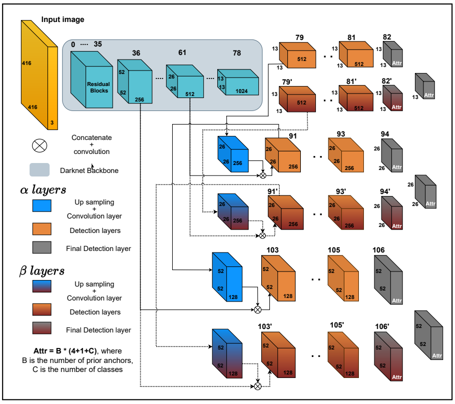


Figure 4: Budding Ensemble Architecture on YOLOv3

robustness. One way to build ensemble DNNs is to train multiple models with different initializations, architectures, or input data and combine their predictions. This paper proposes a novel approach to enhance the reliability of uncertainty quantification in anchor-based object detection networks by introducing redundant layers after the backbone feature extractor. Inspired by the Siamese architecture [3, 5], the proposed approach aims to maximize true positives while minimizing false positives using a differentiable loss function and architecture. This leads to highly calibrated models with improved confidence scores and accurate bounding boxes for detected objects, improving reliability and performance. Detecting out-of-distribution (OOD) images is a crucial task in various machine learning and computer vision applications to ensure model robustness and safety.

Existing OOD detection literature mainly focuses on classification networks [44], employing post-hoc methods like temperature scaling [24], confidence enhancement techniques [8], or training with OOD data [9]. For instance, a multitude of out-of-distribution (OOD) detectors has been developed using generative models such as Variational Autoencoders (VAEs) [36, 43] and Generative Adversarial Networks (GANs) [39]. These detectors typically rely on indicators like reconstruction error or assessing the likelihood of a sample being generated by the model. However, these solutions often come with additional computational overhead, primarily dedicated to OOD detection rather than conventional object detection. In contrast, our proposed BEA-based models offer a dual functionality and minimal overhead. They perform standard object detection seamlessly while concurrently delivering state-of-the-art OOD detection capabilities, thus offering a more efficient and integrated solution.

3 Budding Ensemble Architecture

This section introduces the Budding Ensemble Architecture (BEA), a novel architecture comprising a shared backbone and only two duplicated detectors that is responsible for enhancing the efficiency and effectiveness of this reduced ensemble model compared to traditional object detection ensembles. The BEA not only calibrates prediction confidence scores but also enables the detection of out-of-distribution (OOD) samples. Further, the application of BEA is explained in detail using YOLOv3 model as an example.

3.1 BEA-YOLOv3

Recap on YOLOv3 (base model): In this model, three detectors are placed in the final layers, each dedicated to predicting objects of different scales, similar to the concept of feature pyramid networks (FPN [26]) illustrated in Fig. 3. YOLOv3's final detector layers perceive the input image as ***S x S*** grids, where each grid size S is distinct for each detector layer. The final layer's prediction feature maps are tasked with object detection using three anchors (***B***) with distinct aspect ratios on the corresponding grid of an input image. Therefore, there are three predictions per grid at each scale of the prediction maps in the final detection layers.

The YOLOv3 model is transformed into the BE Architecture shown in Figure 2 and 4 by duplicating the detector layers. This results in six detectors, compared to the three in the base YOLOv3 model. We refer to the original and duplicate layers as **tandem layers**. Assuming that the initial setup of creating the tandem layers is accurate and combining the predictions from these layers is correct, it is reasonable to expect that the model's accuracy and confidence score calibration will enhance. This is because the model can analyze the intermediate features twice by sending the same representations to both detectors, which aids in capturing uncertainty. However, during regular training of the vanilla form of BEA YOLOv3 architecture with its conventional loss functions (using $\mathcal{L}_{\mathbf{conv}}$), both the original detectors (α) and duplicate detectors (β) end up learning similar representations, confidently agreeing on both correct and incorrect predictions. To enhance the model's uncertainties and calibration, we aim for a strong disagreement on incorrect predictions and agreement on correct predictions. To achieve this, we propose two new loss functions: the Tandem-Quelling loss function ($\mathcal{L}_{\mathbf{tq}}$) (eq. 1) to encourage disagreement between negative predictions when no object is present in the grid, and the Tandem-Aiding loss function ($\mathcal{L}_{\mathbf{ta}}$) (eq. 2) to promote agreement between positive predictions when an object is present at the individual output of both the regression (bounding box information x,y,w,h) and classification (C). This setup also allows the model to exchange information when one of the tandem detectors is not confident about its prediction.

$$\mathfrak{L}_{\text{tq}}(\hat{\phi}) = \Sigma_{i=1}^{S^2} \Sigma_{j=1}^{B} \mathbb{1}_{ij}^{\text{noobj}} \frac{2}{\sqrt{\left(\hat{\phi}_i^{\alpha} - \hat{\phi}_i^{\beta}\right)^2}}, \quad \mathcal{L}_{\text{tq}} = \mathfrak{L}_{\text{tq}}(\hat{x}) + \mathfrak{L}_{\text{tq}}(\hat{y}) + \mathfrak{L}_{\text{tq}}(\hat{w}) + \mathfrak{L}_{\text{tq}}(\hat{h}) + \mathfrak{L}_{\text{tq}}(\hat{C}) \tag{1}$$

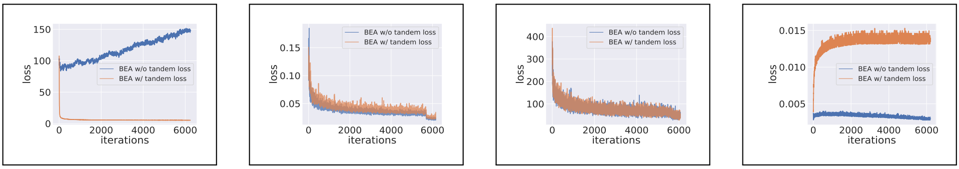
$$\mathfrak{L}_{\text{ta}}(\hat{\phi}) = \Sigma_{i=1}^{S^2} \Sigma_{j=1}^{B} \mathbb{1}_{ij}^{\text{obj}} \frac{\sqrt{\left(\hat{\phi}_i^{\alpha} - \hat{\phi}_i^{\beta}\right)^2}}{2}, \quad \mathcal{L}_{\text{ta}} = \mathfrak{L}_{\text{ta}}(\hat{x}) + \mathfrak{L}_{\text{ta}}(\hat{y}) + \mathfrak{L}_{\text{ta}}(\hat{w}) + \mathfrak{L}_{\text{ta}}(\hat{h}) + \mathfrak{L}_{\text{ta}}(\hat{C}) \tag{2}$$

$$\mathcal{L}_{\text{tandem}} = \mathcal{L}_{\text{tq}} + \mathcal{L}_{\text{ta}} \tag{3}$$

$$\mathcal{L}_{\text{bea}} = \mathcal{L}_{\text{conv}} + \mathcal{L}_{\text{tandem}} \tag{4}$$

where $\mathbb{1}_{ij}$ in eq. 1 and 2 denotes that the object in ground truth appears in j_{th} anchor or the bounding box predictor and grid i is responsible for the particular prediction. $\hat{\phi}$ is a tuple consists of specific outputs from α and β detector, $\hat{\phi} = (\hat{\phi}^{\alpha}, \hat{\phi}^{\beta})$. Collectively the **Tandem loss functions** - $\mathcal{L}_{\mathbf{tandem}}$ (eq. 3) operate on the tandem detectors α and β to increase the variance between their negative predictions. Similarly, they decrease the variances between the positive predictions.

Aggregation of Tandem detectors: During the inference stage, the tandem detectors are combined by averaging their bounding box coordinates, resulting in tighter boxes closer to the ground truth objects. However, only the maximum objectness and class scores from the tandem detectors are carried forward. Before discarding the tandem detectors, the $\mathcal{U}ood$ for OOD detection is captured at each anchor as explained in section 4.1. Aggregated



(a) (b) (c) (d)

Figure 5: Monitoring tandem loss functions ($\mathcal{L}_{\mathbf{tandem}}$) on BEA-YOLOv3. WH indicates predicted width and height offsets of the bounding box: (a) Monitoring tandem-quelling (L_{tq}) on WH; (b) Monitoring tandem-aiding (L_{ta}) on WH; (c) Original WH loss; (d) Normalised MSE between α and β detectors

detector predictions are sent to a Non-maximum suppression (NMS) module, which uses modified bounding box coordinates to remove overlapping entities and lower confidence score predictions. The method involves maximum confidence score voting and averaging of bounding boxes.

Uncertainty induced by Tandem loss functions: Fig. 5 shows how the individual factors of the loss function $\mathcal{L}_{\mathbf{tandem}}$ operate. The addition of $\mathcal{L}_{\mathbf{tandem}}$ to the vanilla BEA form increases the variance between negative predictions, causing the $\mathcal{L}_{\mathbf{tq}}$ loss to decrease, as seen in Fig. 5. However, the introduction of $\mathcal{L}_{\mathbf{tandem}}$ does not affect the $\mathcal{L}_{\mathbf{ta}}$ loss (as shown in Fig. 5), since the original loss functions of YOLOv3 already operate on positive predictions individually for each detector. It is worth noting that the $\mathcal{L}_{\mathbf{ta}}$ loss plays a crucial role in reestablishing confidence in positive predictions with low variance between the tandem layers. The introduced loss function's effectiveness is validated by monitoring the mean squared error (mse) between the α and β detectors (Fig. 5). This approach helps suppress false positives during training and improves the model's confidence score calibration, enhancing trust in the model.

Tradeoff between $\mathcal{L}_{\mathbf{conv}}$ and $\mathcal{L}_{\mathbf{tandem}}$: We conducted experiments by varying the weights assigned to $\mathcal{L}_{\mathrm{conv}}$ and $\mathcal{L}_{\mathrm{tandem}}$ to assess their influence on performance and uncertainty error. Assigning a weight greater than 1 to $\mathcal{L}_{\mathrm{tandem}}$ did not lead to a significant reduction in uncertainty error or an enhancement in OOD detection. Instead, it had an adverse impact on prediction performance compared to the baseline. To achieve a balanced improvement across predictions, reduction in uncertainty error, and enhanced OOD detection, it is imperative to allocate equal weight to both loss components.

3.2 BEA on Gaussian-YOLOv3 and SSD

To demonstrate the proposed method's effectiveness and versatility, we applied it to the Gaussian-YOLOv3 model [6] and SSD [29], as previously discussed in this section. Gaussian-YOLOv3 is a modified version of YOLOv3 where the bounding box coordinates ($\hat{\mathrm{x}}$, $\hat{\mathrm{y}}$, $\hat{\mathrm{w}}$, and $\hat{\mathrm{h}}$) are modelled as Gaussian outputs using negative log-likelihood (NLL) loss functions. Despite this alteration, the functionality of $\mathcal{L}_{\mathbf{tandem}}$ remains unchanged, as it operates on the α and β detectors, whose predictions are generated by the Gaussian function. The BEA technique is also applied to SSD by replicating only the **Detections** layers, as illustrated in Figure 2 of the paper [29]. Being an orthogonal approach, this demonstrates BEA's ability to be applied to different object detection models. The point of split where the detectors are duplicated is an interesting factor which can impact the efficiency of BEA. This is further discussed using SSD as an example in the supplementary material.

4 Evaluation of BEA

Experimental Setup: We compare the performance of the proposed BEA architecture for object detection with different versions of YOLOv3 and SSD models, trained using a standardized approach with identical hyperparameters and an input image size of 416 × 416 using the mmdetection framework [4]. YOLOv3 ensembles were trained for 300 epochs using the bagging method [12] with pre-trained weights from COCO [25]. The evaluation was performed on the KITTI dataset [14] with seven usable classes out of nine, and the dataset was split using a ratio of 7.5:1:1.5 for training, validation, and testing to evaluate the model's generalization ability.

4.1 Evaluation Metrics

In this section, we will discuss the evaluation metrics used to measure the effectiveness of applying BEA to YOLOv3 and SSD. The subsequent analysis delves into the experiments and their results in detail.

Uncertainty Error (UE): We use the uncertainty error metric [32] to assess the model's ability to distinguish correct and incorrect detections based on the uncertainty estimate threshold (δ).

$$TPRj = \frac{|U(\mathbb{D}_c) > \delta|}{|\mathbb{D}_c|}, \quad FPRt = \frac{|U(\mathbb{D}_i) \leq \delta|}{|\mathbb{D}_i|}, \quad UE = \frac{TPRj + FPRt}{2} \tag{5}$$

$$\mathcal{U}_{pred} = 1 - \hat{C} \tag{6}$$

where *TPRj* (True Positive Rejection) is a proportion of correct detections $\mathbb{D}_c$ which are incorrectly rejected (U is the uncertainty measure of detections $\mathbb{D}$), and *FPRt* (False Positive Retention) is a proportion of incorrect detections $\mathbb{D}_i$ which are incorrectly accepted. The optimal value of δ is chosen, giving maximum true positives and minimum false positives out of all combined detections. A perfect uncertainty error score of 0% indicates that all incorrect detections ($\mathbb{D}_i$) are rejected, and all correct detections ($\mathbb{D}c$) are accepted. Both the true positive rate and the false positive rate are equally weighted. Since the BEA incorporates Tandem Loss functions that calibrate the model's confidence score, the predicted object uncertainty $\mathcal{U}_{pred}$ can be inferred directly as the complement of the confidence score. **Detection Accuracy (Average Precision)** We evaluate the accuracy of our object detection models using two variants of the Average Precision (AP) metric: mAP and AP50, which measure the mean AP over different IOU thresholds ([50:5:95]) and the AP at a 50% intersection over the union (IOU) threshold, respectively. We calculate $\mathbf{AP50}_{\mathcal{U}_{\mathbf{raw}}}$ and $\mathcal{U}_{pred}$-based $\mathbf{AP50}_{\mathcal{U}_{\mathbf{pred}}}$ scores to demonstrate the effectiveness of the BEA approach. We compute the raw AP50 by applying post-NMS predictions using a confidence threshold of 0.05. For the $\mathcal{U}_{pred}$-based AP50 ($\mathbf{AP50}_{\mathcal{U}_{\mathbf{pred}}}$), we exclude samples with $\mathcal{U}_{pred} > \delta$. A well-calibrated model with precise uncertainty estimates will likely have AP50 scores closer to the raw AP50.

AP50-based Retention Curves: Inspired by the Shifts benchmark [31], we adopt retention curves, commonly used for regression tasks, to assess the object detection model's robustness and quality of the calibration. Reliability diagrams [7] are unsuitable for object detection models as they are primarily used to measure the calibration of confidence scores in classification tasks. Retention curves are better suited for the complex AP metric involving regression and classification. Specifically, we use AP50-based retention curves, which involve sorting predictions by their descending uncertainty level and calculating AP50 scores at various fractions of retained predictions as illustrated in Fig. 7. The AUC of these curves measures the calibration quality of confidence scores. A poorly calibrated model will need a larger fraction (x-axis) to attain the **AP50**$_{\mathcal{U}_{\mathbf{raw}}}$. This metric should only compare the model's calibration using the same validation dataset to avoid bias.

Retention Curves vs. Uncertainty Error: To comprehensively evaluate object detection models, consider both AP50-based retention curves and uncertainty error metrics. Note that models with fewer detections may have lower UE metric values for *TPRj* or *FPRt* (eq. 5), indirectly reducing the UE metric. Precision-recall curves may not penalize false positives with low confidence scores, affecting AP50-based retention curves [35].

Out of distribution (OOD) detection: The uncertainty measure $\mathcal{U}_{ood}$ is calculated by taking the mean (μ) of the entropy value for each grid cell and anchors from the prediction of both detectors, as shown in eq. 7. This metric determines whether input data is in-distribution or out-of-distribution by quantifying whether out-of-distribution input is correctly detected, i.e., assigned a high uncertainty. In contrast, in-distribution data should be assigned a low uncertainty value. To evaluate the effectiveness of the uncertainty measure $\mathcal{U}_{ood}$ in identifying out-of-distribution data, we compute the AUC-ROC [2] using the $\mathcal{U}_{ood}$ values across a complete dataset that combines in-distribution and out-of-distribution data in a 1:2 ratio. Using the BEA property with $\mathcal{L}_{\mathbf{tandem}}$, we combine the mean squared error of tandem bounding box prediction (x̂, ŷ, ŵ, and ĥ), confidence score, and entropy from α and β detectors as an uncertainty measure $\mathcal{U}_{ood}$, as shown in eq.7 and eq.8. The (s,b) in eq.7 and eq.8 refers to all anchors which are accessed individually.

$$\mathfrak{B}(z,i,j) = \left(\hat{z}^{\alpha}_{ij} - \hat{z}^{\beta}_{ij}\right)^2, \quad \mathcal{B}(s,b) = \sqrt{\sum_{z\in(\hat{x},\hat{y},\hat{w},\hat{h},\hat{C})} \mathfrak{B}(z,s,b)}, \quad H(X) = -\sum_{i=1}^{n} p(x_i)\ln(p(x_i)),$$
$$\mathfrak{h}(z,i,j) = \left(H(\hat{z}^{\alpha}_{ij}) - H(\hat{z}^{\beta}_{ij})\right)^2, \quad \mathcal{H}(s,b) = \sqrt{\sum_{z\in\hat{c}} \mathfrak{h}(z,s,b)}. \tag{7}$$

$$\mathcal{U}_{ood} = \mu\Big(\mathcal{B}(s,b) * \mathcal{H}(s,b)\Big) \tag{8}$$

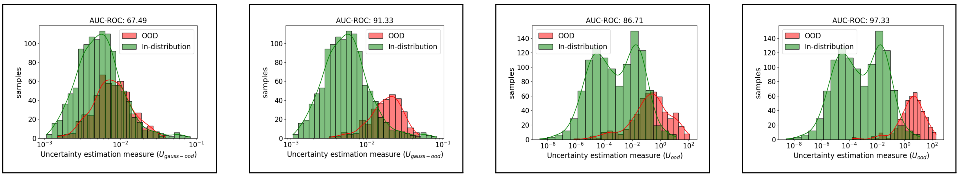
For non-BEA-YOLOv3 and non-Gaussian-based YOLOv3 models, we utilize their confidence scores as their $\mathcal{U}_{ood}$. However, for Gaussian-based YOLOv3 models, we use its $Uncertainty_{aver}$ as $\mathcal{U}_{ood}$. To assess the OOD detection performance of the models, we use two near-OOD datasets $\mathcal{U}_{near-ood}$ (CityPersons [47], and BDD100K [45]) and one far-OOD dataset $\mathcal{U}_{far-ood}$ (COCO [25]) and compute the AUC-ROC values. Higher AUC-ROC values indicate better OOD detection performance.

4.2 Results and Discussion

BEA improves both YOLOv3 and SSD model's calibration and prediction accuracy with low uncertainty error (UE), enabling it to detect out-of-distribution data. The model's accuracy is evaluated using the average precision metric, which requires a minimum overlap of 50%, but a minimum confidence score of 0.05 can lead to higher false positives. An accurate uncertainty estimation model can identify and discard samples with high uncertainty scores, resulting in lower UE. This section discusses Table 1, particularly the YOLOv3 results in detail. Table 1 shows that YOLOV3 with BEA has the lowest UE compared to Base-YOLOv3, while Gaussian-YOLOv3 has a better UE (around 4.9%). Incorporating BEA into Gaussian-YOLOv3 further reduces UE to approximately 4%, demonstrating BEA's significant improvement in uncertainty estimation, regardless of the detection model's conventional loss functions.

Models (input size 416×416)	**mAP$_{raw}$** (%) ↑	**AP50** ↑		**UE** (%) ↓	**AP50-based Retention curve AUC** (%) ↑	**Out-of-distribution detection (OOD) AUC-ROC** (%)↑		
		AP50$_{raw}$	**AP50$_{\mathcal{U}_{pred}}$**			**CityPersons** $\mathcal{U}_{near-ood}$	**BDD100K** $\mathcal{U}_{near-ood}$	**COCO** $\mathcal{U}_{far-ood}$
Base-YOLOv3	51.72	87.4	78.2	11.96	53.1	35*	40.16*	20.21
YOLOv3 3 Ensemble	54.58	89	82.94	9.23	58.7	28.79*	32.44*	20.5*
YOLOv3 5 Ensemble	55.1	89.27	82.97	9.03	59.3	28.6*	12.19*	10.21*
BEA-YOLOv3	**54.83 ± 0.28**	**89.3 ± 0.28**	**85.79 ± 0.13**	**4.55 ± 0.02**	**73.9 ± 1.1**	**98.75 ± 2.3**	**86.71 ± 1.7**	**97.33 ± 0.9**
Gaussian YOLOv3	47.65	88.17	83.98	4.96	81.8	78.98**	67.49**	91.33**
Gaussian YOLOv3 + 3 Ensemble	50.43	89.48	85.66	4.61	84.4	81.72**	71.08**	89**
Gaussian YOLOv3 + 5 Ensemble	52.29	89.92	85.79	4.55	84.7	82.31**	71.56**	84.8**
BEA-Gaussian-YOLOv3	**54.28 ± 0.14**	**90.64 ± 0.34**	**86.5 ± 0.1**	**4.05 ± 0.01**	**86.2 ± 0.4**	**79.21 ± 1.7**	**85.89 ± 3**	**98.4 ± 1.1**
Base-SSD	51.24	88.69	80.15	11.28	73.5	42.95*	45.41*	26.37*
SSD + 5 Ensemble	52.8	89.14	82.71	10.06	78.6	38*	41.28*	37.8*
BEA-SSD	**53.18 ± 0.08**	**90.38 ± 0.17**	**86.83 ± 0.4**	**7.7 ± 0.08**	**82.5 ± 0.4**	**61.3 ± 2.1**	**61.38 ± 3.4**	**88.49 ± 0.87**

Table 1: Evaluation of the BEA architecture on YOLOv3 and SSD using various metrics, including accuracy, uncertainty estimation, calibration, robustness, and retention curves. Trained on all seven usable classes of KITTI with a constant set of hyperparameters, the test set comprised 15% of KITTI data. The effectiveness of OOD detection was demonstrated using one far-ood and two near-ood datasets. * represents that the confidence score is used for detecting OOD data, and ** represents Gaussian uncertainty is used for the same purpose.



(a) (b) (c) (d)

Figure 6: Out-of-distribution image detection with KITTI-trained YOLOv3 models on COCO and BDD1000K datasets: (a) Gauss-YOLOv3 - BDD100K; (b) Gauss-YOLOv3 - COCO; (c) BEA-YOLOv3 - BDD100K; (d) BEA-YOLOv3 - COCO

BEA improves YOLOv3 accuracy with mAP_{raw} and $AP50_{raw}$ increasing by 6% and 3.7% respectively over base-YOLOv3. BEA-Gaussian-YOLOv3 shows 14% and 2.8% improvement of mAP_{raw} and $AP50_{raw}$ over Gaussian-YOLOv3. The Gaussian-modeled loss function used in Gaussian YOLOv3 enables it to excel at learning from more evidence and perform better on dominant classes. However, Gaussian YOLOv3 may not perform well for less prominent classes, leading to a lower-than-expected mAP_{raw} when evaluated with all seven classes of the KITTI dataset. Discarding predictions with high uncertainty scores using $\mathcal{U}pred$ has little impact on BEA's $AP50\mathcal{U}_{pred}$, indicating more accurate uncertainty estimates. BEA achieves approximately 9.6% higher AP50 than base-YOLOv3 and 4.2% higher AP50 than Gaussian-YOLOv3 using these uncertainties. The uncertainty measure $AP50_{\mathcal{U}_{pred}}$ considers the overall uncertainty of a prediction, as the Tandem Loss functions work collectively to reduce variance using all factors of object prediction. Our uncertainty measure doesn't separate uncertainty into spatial or localization uncertainty. Our Gaussian-YOLOv3 implementation has a lower uncertainty error than Gasperini et al. [13].

Fig. 7 displays the AP50 Retention curves for Base-YOLOv3, 5-member ensemble, Gaussian-YOLOv3, and BEA models. The findings reveal that BEA-YOLOv3 performs better than Base-YOLOv3, with a 40% increase in the area under the AP50-based retention curve. Additionally, BEA Gaussian-YOLOv3 shows a 5.4% enhancement over Gaussian-

YOLOv3, demonstrating improved calibration of confidence scores. BEA's effectiveness is shown to depend on the base loss functions and their ability to train α and β detectors to learn representations independently, as observed in the significant improvement in the UE results when BEA is incorporated into the Gaussian version. Additionally, BEA's aggregation approach, as discussed in Section 3, contributes to its superior performance. The **Tandem Loss** functions improve confident predictions at each tandem pair during training. Disagreements in negative predictions between bounding boxes, confidence scores, and objectness scores can identify out-of-distribution (OOD) data when detectors disagree with their corresponding tandem detector β.

BEA's OOD detection ability is evaluated by combining in-distribution with out-of-distribution datasets at a 2:1 ratio of 1000 and 500 validation images. The BEA model performs better than all other models in OOD detection ability, as shown in Table 1 and Fig. 6. AUC-ROC is calculated by averaging the OOD uncertainty scores $\mathcal{U}_{ood}$ for all predictions made on each image. The model is evaluated with two near-ood datasets (Citypersons [47] and BDD100K [45]) and the in-distribution dataset.

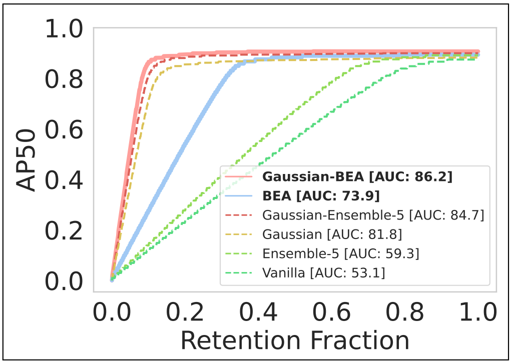


Figure 7: AP50-based retention curves

The BEA model exhibits superior performance compared to all other models, with a significant margin of improvement. BEA has a lower FPS than Base-YOLOv3 due to sequential execution, but parallelizing tandem layers could improve this. Despite having fewer parameters than ensembles, BEA surpasses them in accuracy, calibration, and uncertainty estimates.

5 Conclusion and Future Work

In conclusion, this study proposes a novel Budding-Ensemble Architecture (BEA) for Single-Stage Anchor-based models, which includes a new loss function called tandem loss functions. The proposed architecture enhances detection accuracy and well-calibrated confidence scores leading to higher-quality uncertainty estimates. Evaluation results using various metrics demonstrate the superiority of BEA over base models (YOLOv3 and SSD) and ensembles regarding calibration and uncertainty estimation. The architecture's unique structural advantage provides an opportunity to enhance model reliability in different settings. Future work can optimize and integrate BEA into vision-based models, potentially replacing ensembles for improved accuracy and efficiency. Additionally, future work will study the architecture split point where detectors are duplicated to understand its impact on score calibration and OOD detection. Additionally, BEA's effectiveness can be explored in more complex scenarios, such as object detection in crowded or cluttered scenes. Overall, BEA is a promising framework for improving object detection model reliability and optimizing its integration can lead to significant improvements in model performance on various tasks.

Acknowledgments: This work was partially funded by the Federal Ministry for Economic Affairs and Climate Action of Germany, as part of the research project SafeWahr (Grant Number: 19A21026C)

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[2] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Davide Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021.

[6] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 502–511, 2019.

[7] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2): 12–22, 1983.

[8] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[9] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

[10] Michael W. Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-An Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. 2020.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[13] Stefano Gasperini, Jan Haug, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, Benjamin Busam, and Federico Tombari. Certainnet: Sampling-free uncertainty estimation for object detection. *IEEE Robotics and Automation Letters*, 7(2):698–705, 2021.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[19] Neslihan Kose, Ranganath Krishnan, Akash Dhamasia, Omesh Tickoo, and Michael Paulitsch. Reliable multimodal trajectory prediction via error aligned uncertainty optimization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 443–458. Springer, 2023.

[20] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020.

[21] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *Proceedings of the 35th International Conference on Machine Learning*, 80:2796–2804, 2018.

[22] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[28] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[30] Chunwei Ma, Ziyun Huang, Jiayi Xian, Mingchen Gao, and Jinhui Xu. Improving uncertainty calibration of deep neural networks via truth discovery and geometric optimization. *Uncertainty in Artificial Intelligence (UAI)*, pages 75–85, 2021.

[31] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

[32] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 international conference on robotics and automation (icra)*, pages 2348–2354. IEEE, 2019.

[33] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[34] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74, 1999.

[35] Syed Qutub, Florian Geissler, Yang Peng, Ralf Gräfe, Michael Paulitsch, Gereon Hinz, and Alois Knoll. Hardware faults that matter: Understanding and estimating the safety impact of hardware faults on object detection dnns. In *Computer Safety, Reliability, and Security: 41st International Conference, SAFECOMP 2022, Munich, Germany, September 6–9, 2022, Proceedings*, pages 298–318. Springer, 2022.

[36] Xuming Ran, Mingkun Xu, Lingrui Mei, Qi Xu, and Quanying Liu. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *Neural Networks*, 145:199–208, 2022.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[39] Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, 2018.

[40] Marius Schubert, Karsten Kahl, and Matthias Rottmann. Metadetect: Uncertainty quantification and prediction quality estimates for object detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2021.

[41] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 119:9690–9700, 13–18 Jul 2020.

[42] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[43] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33:20685–20696, 2020.

[44] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[45] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[46] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *International Conference on Learning Representations*, 2020.

[47] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017.